Biclustering Analysis and Comparison for Gene Expression Data

Lu Yin\textsuperscript{1,2,3}, Yongguo Liu\textsuperscript{1,3,*}

\textsuperscript{1} School of Computer Science and Engineering University of Electronic Science and Technology of China ChengDu, SiChuan, 611731, P. R. China; \\
\textsuperscript{2} School of Computer Engineering Huaiyin Institute of Technology HuaiAn, JiangSu, 223003, P. R. China; \\
\textsuperscript{3} School of Information and Software Engineering University of Electronic Science and Technology of China ChengDu, SiChuan, 610054, P. R. China; \\

*Corresponding author: E-mail: liuyg@uestc.edu.cn

Abstract

The advance of the high throughput technology allows to quantitative measure the level of expression of thousands of genes under several experimental conditions at one time which quickly produce many gene expression data. For efficiency processing the above data, many biclustering algorithms and many criterions are proposed. Even so, different biclustering algorithms generate different type of biclusters and draw various conclusions because the definition of the bicluster and its evaluation criterions used indifferent biclustering algorithms are similar but not the same. But there are only a few literatures for comparing the biclustering algorithms and its evaluation criterions. In this paper, the biclusters obtained by 11 biclustering algorithms are evaluated according to 7 criterions. In addition, the weight enrichment score and the proportion of the enriched biclusters by gene ontology are used for evaluating the biological significance of the biclustering algorithms. By analyzing the results of the chose biclustering algorithms, this study suggest that the biclustering algorithms provide a flexible solution for analyzing multiple types of gene expression data, and many interesting results are observed which facilitate exploratory data mining of gene expression data.

Key words: biclustering, gene expression data, gene ontology, gene set enrichment analysis.

1. INTRODUCTION

With the development of the high throughput genomic technology, researchers can simultaneously monitor thousands of genes and explore a functional genome invarious samples. As a result, the amount of gene expression data greatly increases, which necessitates an efficient method to analyze and process these data in order to find
similar or relevant gene clusters, conjecture the function of the unknown genes, and explore the gene regulatory network (Jiang et al., 2004). For example, 8600 human genes were clustered and discussed how to identify the gene function (Eisen et al., 1998). The gene regulatory network was identified through using k-mean algorithm in yeast mRNA expression data (Tavazoie et al., 1999). Although the abovementioned standard clustering methods are able to analyze gene expression data (Tanay et al., 2002), research on cancer and gene knockout has proved that it is difficult for these methods to identify local gene expression patterns where some gene functions are only correlated to some samples (Getz et al., 2000; Alizadeh et al., 2000). Additionally, in the standard clustering method, each gene must be divided in one cluster, but biological experiments have demonstrated that some genes take part in one or more biological functions, and in some cases even no function (Ihmels et al., 2002; Gasch and Eisen, 2002). The fuzzy k-means clustering was adopted to analyze the conditional co-regulation on the yeast stress condition expression data which found that many genes can be assigned to multiple clusters for revealing distinct aspects of their functions (Gasch and Eisen, 2002). So the standard clustering algorithm cannot accurately analyze the gene expression data and discover the local gene expression patterns that are significant for the researchers.

In order to solve the shortcomings of the standard clustering method, some techniques were redesigned (Getz et al., 2000; Alizadeh et al., 2000; Ihmels et al., 2002; Gasch and Eisen, 2002; Tamayo et al., 1999; Cheng and Church, 2000; Lazzeroni and Owen, 2002; Ben-Dor et al., 2002; Murali and Kasif, 2003). Among the above method, the biclustering has become essential in analyzing and processing the gene expression data. The CC biclustering algorithm was firstly proposed to handle the gene expression data which took Mean Square Residue (MSR) as the evaluating criterion and reduced MSR to the given threshold by removing or adding rows or columns (Cheng and Church, 2000). The problem of biclustering gene expression data has proved to be NP-hard (Cheng and Church, 2000; Hartigan, 1972). The biclustering algorithms were classified into two categories: systematic search algorithm and stochastic search algorithm. The systematic search algorithm can be further divided into three categories (Ayadi et al., 2009; Maderia and Oliverial, 2004): greedy iterative algorithm, such as ISA (Ihmels et al., 2002), CC (Cheng and Church, 2000), Plaid (Lazzeroni and Owen, 2002), OPSM (Ben-Dor et al., 2002), xMotf (Murali and Kasif, 2003), FLOC (Yang et al., 2002; Yang et al., 2003), BioNMF (Pedro et al., 2006), divide and conquer algorithm, such as Direct clustering (Hartigan, 1972), Bimax (Prelić et al., 2006) and enumeration algorithm, such as SAMBA (Tanay et al., 2002), BiMine (Ayadi et al., 2009) and OP-Cluster (Liu and Wang, 2003). The stochastic searching algorithms are those that adopt a meta-heuristic strategy to explore biclusters, such as SEBI (Divina and Aguilar-Ruiz, 2006), SSB (Nepomuceno et al., 2011), CMOPSOb (Liu and Li, 2009), BIC-aNNet (de Franca et al., 2006) and SAB (Bryan et al., 2006). There are many detailed survey about the biclustering algorithm (Maderia and Oliverial, 2004; Busygine et al., 2008; Mukhopadhyay et al., 2010).

Although many biclustering algorithms have been proposed to identify the bicluster, different algorithms generate different biclusters and draw distinct conclusions because the definition and type of the bicluster provided by various algorithms are similar but not the
same. So comparing biclustering algorithms and analyzing their performance is a challenge. The performance of CC, ISA, OPSM, SAMBA, xMotif, and Bimax was compared on two data according to the proportion of the enriched biclusters by GO, and found that the scores of CC and xMotif are significantly lower than those of other biclustering algorithms (Prelić et al., 2006). 21 clustering algorithms and 6 biclustering algorithms (CC, ISA, OPSM, xMotif, Bimax, and bioNMF) were compared on five gene expression data according to the number of significantly enriched GO terms of the three most enriched biclusters and the p-value of significantly enriched GO terms of the most enriched bicluster, and found that SAMBA has outperformed the others (Bhattacharya et al., 2012). The biclustering algorithms of Bimax, FABIA, ISA, QUBIC and SAMBA was compared on two gene expression data according to the WE score and the PPI score and found that the performance of the biclustering algorithm is closely dependent on the data (Li et al., 2012). The biclustering algorithms of CC, OPSM, SAMBA, MSSRCC, and CPB were compared on three gene expression data according to the p-value of significantly enriched GO terms of the 10 most enriched biclusters and determined that CPB and MSSRCC can identify the more significantly enriched biclusters (Bozdağ et al., 2010). The biclustering algorithms of CC, Plaid, Bimax, OPSM, ISA, QUBIC, BBC, COALESCE, CPB, FABIA, Spectral and xMotif were compared on eight gene expression data according to the proportion of the enriched biclusters by GO, and found that CPB and BBC can identify the biclusters that were most enriched and Plaid can provide the highest proportion of enriched biclusters (Eren et al., 2012). The biclustering algorithms of CC, ISA, OPSM, and SAMBA were compared to six gene expression data in terms of the differential co-expression scoring function and found that the biclustering algorithms are dependent on the expression data and the bicluster type (Chia and Karuturi, 2010). The FLOC, CTWC, Spectral, bioNMF, Plaid, BBC, GBC, SAMBA, Bimax, BiHEA were compared on the BicatYeast data in terms of the proportion of the enriched biclusters by GO (Zhao et al., 2012). These research works are vital for comparing different biclustering algorithms for different expression data.

However, the stochastic searching algorithms receive less attention in the abovementioned comparing projects. For comparing the effectiveness of the stochastic searching algorithms as well as the systematic searching algorithms in exploring the biclusters and for understanding the potential influence of the various criteria combined with the stochastic searching algorithms, six systematic searching algorithms (CC, FLOC, ISA, OPSM, SAMBA and Bimax) and five stochastic searching algorithms (SEBI, BIC-aINet, SAB, SSB and PSOB) under four criteria (MSR, VE (Divina et al., 2012), ACV (Nepormuceno et al., 2011), and ASR (Ayadi et al., 2009)) are compared in order to compare their performance and biology significance.

The rest of the article is organized as follows. Section 2 mathematically describes the type and the evaluating criterion of a bicluster. Section 3 discusses the compared biclustering algorithms and its parameter selection and briefly describes the expression data used in our experiment. In section 4 the aggregated evaluating criterion on three expression data are reported and its biological significance are also discussed. Section 5 lists the most significant results obtained by the above analysis.
2. THE BICLUSTERING PROBLEM

2.1 Concepts of bicluster

After preprocessing a gene expression data is an n×m matrix A(X, Y), in which X = \{x_1, x_2, \ldots, x_n\} is the genes set, Y = \{y_1, y_2, \ldots, y_m\} the samples set, and the element a_{ij} reflects the expression value of the ith gene under the jth sample. Let I and J are the gene subset of X and the sample subset of Y, this is, I \subseteq X, |I| = k, k \leq n) and J \subseteq Y, |J| = l, l \leq m), in which k = |I| and l = |J| (Maderia and Oliverial, 2004).

Definition 1: A cluster of genes is a subset I including similar gene expression patterns across all samples and denoted as k × m submatrix G(I, Y).

Definition 2: A cluster of samples is a sample subset J including similar sample expression profiles across all genes and denoted as n × l submatrix C(X, J).

Definition 3: A bicluster is a k × l sub-matrix B(I, J) consisting of I and J, in which the gene expression pattern of I under J shows similar.

A bicluster has four types (1) bicluster with constant values, (2) bicluster with constant rows or columns, (3) bicluster with a shifting or scaling pattern, and (4) bicluster with coherent evolutions.

Definition 4: A bicluster with constant values means that elements a_{ij} = c, i \in I, j \in J and c are a constant.

Definition 5: A bicluster with constant values of rows or columns means that the element of each row or column is a constant; i.e., a_{ij} = c + \alpha_i for a_{ij} = c + \beta_j, i \in I, j \in J, \alpha_i is a variable only associated with the row, \beta_j is a variable only associated with the column.

Definition 6: A bicluster with a shifting or scaling model means that the element of each row and each column has shifting or scaling factors; i.e., a_{ij} = c \times \alpha_i or a_{ij} = c \times \beta_j, i \in I, j \in J.

Definition 7: A bicluster with coherent evolutions means that gene expression patterns or sample expression profiles behave similarly.

2.2 Evaluation criterion

Given a bicluster B(I, J), the mean a_{ij} of the i th gene expression pattern, the mean a_{ij} of the j th sample expression profile, and the mean a_{ij} of the bicluster B(I, J) are presented respectively as follows:

\[ a_{ij} = \sum_{i \in I} a_{ij} / |I| \] (1)

\[ a_{ij} = \sum_{j \in J} a_{ij} / |J| \] (2)
\[ a_{ij} = \sum_{i \in I, j \in J} a_{ij} / |I||J| \]  

(3)

As a result, the gene number \( r_{ij} \), the sample number \( s_{ij} \), the volume \( v_{ij} \), the variance \( V A R_{ij} \), the row variance \( r V A R_{ij} \), and the column variance \( c V A R_{ij} \) are given respectively as follows:

\[ r_{ij} = |I| = k \]  

(4)

\[ s_{ij} = |J| = l \]  

(5)

\[ v_{ij} = |I| \times |J| = k \times l \]  

(6)

\[ V A R_{ij} = \sum_{i \in I, j \in J} (a_{ij} - a_{ij})^2 / |I||J| \]  

(7)

\[ r V A R_{ij} = \sum_{i \in I, j \in J} (a_{ij} - a_{ij})^2 / |I||J| \]  

(8)

\[ c V A R_{ij} = \sum_{i \in I, j \in J} (a_{ij} - a_{ij})^2 / |I||J| \]  

(9)

The variance \( V A R_{ij} \) was used to identify the bicluster in direct cluster (Hartigan, 1972), and it was adopted to find the biclusters with constant values (Cho et al., 2004).

Definition 8: Mean Squared Residue (MSR) of a bicluster \( B(I,J) \) is defined as:

\[ MSR_{ij} = \sum_{i \in I, j \in J} r_{ij}^2 / |I||J| \]  

(10)

In (10) \( r_{ij} \) denotes the residue of \( a_{ij} \) and it is defined as:

\[ r_{ij} = a_{ij} - a_{ij} - a_{ij} + a_{ij} \]  

(11)

MSR is firstly proposed to explore the bicluster with constant values (Cheng and Church, 2000). It is also employed other biclustering methods such as FLOC, SEBI, and SAB. However, MSR is only suitable to identify the bicluster with constant values and the one with shifting model (Divina et al., 2012; Aguilar-Ruiz et al., 2005).

Definition 9: Virtual pattern of a bicluster \( B(I,J) \) is a row vector \( \rho = \{\rho_1, \ldots, \rho_i, \ldots, \rho_l\} \) and \( \rho_j = a_{ij} \).

Definition 10: Virtual Error (VE) of a bicluster \( B(I,J) \) is defined as:

\[ V E_{ij} = \sum_{i \in I, j \in J} (\hat{a}_{ij} - \hat{\rho}_j)^2 / |I||J| \]  

(12)

In (12) \( \hat{a}_{ij} = (a_{ij} - a_{ij}) / std(a_{ij}) \) and \( \hat{\rho}_j = (\rho_j - \hat{\rho}) / \sigma(\rho) \).

Considering that MSR is unsuitable for the bicluster with scaling model and coherent evolutions, VE was proposed to evaluate the quality of the bicluster (Divina et al., 2012; Pontes et al., 2015). Since VE denotes the difference between the gene expression pattern and virtual pattern, the more similar the gene expression patterns of the bicluster, the smaller its VE.
Definition 11: Average Correlation Value (ACV) of a bicluster $B(I,J)$ is defined as:

$$ACV_{ij} = \max\left\{ \frac{\sum_{i \in I \land j \neq j} r_{ij} / k(k - 1)}{\sum_{i \in I \land j \neq j} r'_{ij} / l(l - 1)} \right\}$$

(13)

In (13) $r_{ij}$ or $r'_{ij}$ is the Pearson correlation coefficient between the $i$th and $j$th row or column vector of $B(I,J)$. The closer to 1 ACV is, the more correlated the gene expression patterns and the sample expression profiles (Teng and Chan, 2008). ACV is employed to search the bicluster with coherent evolutions which show that ACV is sensitive to noise and miss value (Nepomuceno et al., 2011; Teng and Chan, 2008; Cheng et al., 2008).

Definition 12: Average Spearman’s Rho (ASR) of a bicluster $B(I,J)$ is defined as:

$$ASR_{ij} = \max\left\{ \frac{\sum_{i \in I \land j \neq j} \rho_{ij} / k(k - 1)}{\sum_{i \in I \land j \neq j} \rho'_{ij} / l(l - 1)} \right\}$$

(14)

In (14) where $\rho_{ij}$ or $\rho'_{ij}$ is the Spearman’s rank correlation coefficient between the $i$th and $j$th row or column vector of $B(I,J)$. It was proved that ASR is robust with an abnormal value and its calculation does not need to normalize the data (Ayadi et al., 2012).

3. COMPARED BICLUSTERING ALGORITHMS AND GENE EXPRESSION DATA

In order to explore the performance of different kinds of biclustering algorithms, 12 biclustering algorithms including six systematic search algorithms and six stochastic search algorithms are chosen. Six systematic search algorithms include four greedy iterative algorithms (CC, FLOC, OPSM and ISA), a divide and conquer algorithm (Bimax), and an enumeration algorithm (SAMBA). Among them, the result of OPSM and Bimax are obtained by the BiCAT toolbox (Barkow et al., 2006) and that of SAMBA by the EXPANDER (Shamir et al., 2006). The other algorithms are implemented in MATLAB referring to their corresponding references. The setting of parameters can be found in their corresponding references. The single-objective edition of the CMOPSOB was implemented and renamed as PSOB. In this experiment, the population size is set to be 100 for all population-based methods. Table 1 summarizes the chosen biclustering algorithms in which the first column shows the name of the chosen biclustering algorithm, the second column shows the evaluation criteria of a bicluster produced by the corresponding algorithm and the third column shows the recommend parameter used by the corresponding algorithm.

Three gene expression data are adopted to compare experimental algorithms: yeast cell cycle expression data (Yeast Cycle) (Cho et al., 1998), diffuse large B-cell lymphoma expression data (DLBCL) (Alizadeh et al., 2000), and yeast stress condition expression data (Gasch Yeast) (Gasch et al., 2000) which are portrayed in Table 2. In addition, the stopping thresholds of four evaluation criteria for three gene expression data are also presented in Table 2. In this work, the biclustering algorithms are implemented by Matlab 2012 and run on the PC with an Intel Core i3-2120 processor and 8 GB real memory. Each algorithm runs 100 independent trials and gets 100 biclusters except Bimax, OPSM, and SAMBA.
Table 1 The settings of parameters of experimental algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Evaluating criterion</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>MSR</td>
<td>$\alpha = 1.2$</td>
</tr>
<tr>
<td>FLOC</td>
<td>MSR</td>
<td>$k = 100, \rho = 0.35$</td>
</tr>
<tr>
<td>ISA</td>
<td>-</td>
<td>$t_G = 3, t_C = 2$</td>
</tr>
<tr>
<td>OPSM</td>
<td>-</td>
<td>Default</td>
</tr>
<tr>
<td>Bimax</td>
<td>-</td>
<td>Default</td>
</tr>
<tr>
<td>SAMBA</td>
<td>-</td>
<td>Default</td>
</tr>
<tr>
<td>SEBI</td>
<td>MSR, VE, ACV, ASR</td>
<td>$p_c = 0.85, p_m = 0.2, p_e = 0.9$</td>
</tr>
<tr>
<td>BIC-aINet</td>
<td>MSR, VE, ACV, ASR</td>
<td>$\epsilon = 0.8$</td>
</tr>
<tr>
<td>SAB</td>
<td>MSR, VE, ACV, ASR</td>
<td>$t_0 = 1000, t_{min} = 200, rate = 0.1, a_{count} = 100, a_{success} = 50$</td>
</tr>
<tr>
<td>SSB</td>
<td>MSR, VE, ACV, ASR</td>
<td>$M_1 = 10, M_2 = 10$</td>
</tr>
<tr>
<td>PSOB</td>
<td>MSR, VE, ACV, ASR</td>
<td>$c_1 = c_2 = c_3 = 1, w = 0.5$</td>
</tr>
</tbody>
</table>

Table 2 Experimental gene expression data and its thresholds of the coherence criterion

<table>
<thead>
<tr>
<th>Data</th>
<th>Gene number</th>
<th>Sample number</th>
<th>MSR</th>
<th>VE</th>
<th>ACV</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast Cycle</td>
<td>2884</td>
<td>17</td>
<td>300.00</td>
<td>0.30</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>DLBCL</td>
<td>4026</td>
<td>96</td>
<td>1200.00</td>
<td>0.30</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Gasch Yeast</td>
<td>2993</td>
<td>173</td>
<td>0.05</td>
<td>0.30</td>
<td>0.80</td>
<td>0.80</td>
</tr>
</tbody>
</table>

4. EXPERIMENT EVALUATING

In our comparing work, the evaluation index have the gene number $r_{ij}$, sample number $s_{ij}$, mean squared residue $MSR_{ij}$, virtual error $VE_{ij}$, Average Correlation Value $ACV_{ij}$, Average Spearman’s Rho $ASR_{ij}$ and runtime of a bicluster $B(i, J)$ in which $r_{ij}$ and $s_{ij}$ jointly reflect the size of $B(i, J)$, $MSR_{ij}$, $VE_{ij}$, $ACV_{ij}$ and $ASR_{ij}$ are the four internal validity indices which measure the quality of $B(i, J)$. Runtime reflects the time performance of each algorithm. For MSR, VE, and time of the above seven criteria, the smaller its value the better the bicluster or the biclustering algorithm, while for the other criteria the higher the better. In addition, since the biclusters of Bimax, OPSM and SAMBA are produced by the BiCAT and EXPANDER and its runtime is not clear, the following time analysis does not involve in Bimax, OPSM and SAMBA. For convenience, the experiments results are compared in terms of the following aspects: volume, coherence criterion, and time.

The number of extracted biclusters for each algorithm implemented by our is set to 100 on each data, however, the number of biclusters of OPSM, SAMBA and Bimax produced by the toolbox of BiCAT and EXPANDER respectively are 13, 7 and 4 on Yeast Cycle data, of 10, 121 and 23 of DLBCL data, of 13, 132 and 142 on GASCH Yeast data. So the number of aggregated biclusters for each algorithm is 300 on three data, except that the number of OPSM is 36, of SAMBA is 260 and of Bimax is 269. These result are aggregated on three data and are presented in Table 3, and its first column presents the name of the chose
algorithms in which include six systematic algorithm and stochastic searching algorithm using MSR, VE, ACV and ASR, the next seven columns respectively show the mean of gene number, sample number, MSR, VE, ACV, ASR and time. For five stochastic searching algorithms these result are also aggregated on four criteria and three data and presented in Table 4, its first column shows the name of the stochastic searching algorithm and the next seven columns are the same as in Table 3.

4.1 Comparison of evaluating criteria

As shown in Table 3, the range of gene number obtained by systematic algorithm is very great, this is, FLOC of the largest gene number and Bimax of the smallest are systematic searching, and the gene number of stochastic algorithm using MSR is far higher than that using VE, ACV and ASR. This clearly demonstrates that for this data the widely used MSR threshold 300 is relatively relaxed (Divina et al., 2012). The sample number of systematic algorithm is higher than that of systematic algorithm in which of Bimax only is 4.65. The mean time of the systematic algorithms is far smaller than that of stochastic algorithms in which the time of using MSR is best, VE is second best being better than ACV and ASR, and using ASR undoubtedly had the worst time performance, this is mainly because the computing complex of four evaluating criteria are different.

As the coherence criterion, the MSR and VE of Bimax is the best which is mainly because that its gene and sample number is the smallest. For stochastic algorithm, it is no surprise that the MSR of the stochastic algorithms using MSR as the evaluating criterion exists the bias such that it is the best. This is equally true of VE, ACV and ASR. So, without regard to the above four cases (which are labeled in italics), the coherence criterion of stochastic searching algorithm using VE as the evaluating criterion is best, being that its mean MSR, VE and ACV is best, and its ASR is second best, while that using MSR as the evaluating criterion is worst.

Table 4 shows that in six stochastic searching algorithms the gene number of BIC-aiNet is the best while the SSB is the worst. Unlike the gene number, the sample number of SSB is the best, and the time performance and four criteria of SSB are the best.

Table 3 The aggregated result of the systematic and stochastic algorithm on three data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Gene</th>
<th>Sample</th>
<th>MSR</th>
<th>VE</th>
<th>ACV</th>
<th>ASR</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>systematic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>147.91±55.82</td>
<td>51.68±12.33</td>
<td>411.33±176.85</td>
<td>0.76±0.01</td>
<td>0.26±0.0</td>
<td>0.24±0.0</td>
<td>0.04±0.0</td>
</tr>
<tr>
<td>FLOC</td>
<td>907.99±126.62</td>
<td>25.62±2.83</td>
<td>499.79±208.11</td>
<td>0.93±0.03</td>
<td>0.11±0.0</td>
<td>0.09±0.0</td>
<td>11.06±2.02</td>
</tr>
<tr>
<td>ISA</td>
<td>229.77±51.86</td>
<td>24.46±6.20</td>
<td>9137.38±957.26</td>
<td>0.58±0.12</td>
<td>0.15±0.0</td>
<td>0.18±0.0</td>
<td>0.06±0.0</td>
</tr>
<tr>
<td>OPSM</td>
<td>401.31±42.59</td>
<td>8.45±0.28</td>
<td>2727.23±174.62</td>
<td>0.51±0.09</td>
<td>0.92±0.0</td>
<td>0.92±0.0</td>
<td>--</td>
</tr>
</tbody>
</table>
Table 4 The aggregated results of the stochastic algorithm on three data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Gene</th>
<th>Sample</th>
<th>MSR</th>
<th>VE</th>
<th>ACV</th>
<th>ASR</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAB</td>
<td>397.42±23.08</td>
<td>47.14±18.44</td>
<td>3481.02±207.85</td>
<td>0.48±0.00</td>
<td>0.61±0.00</td>
<td>0.52±0.00</td>
<td>138.58±0.28</td>
</tr>
<tr>
<td>PSOB</td>
<td>317.01±34.36</td>
<td>61.86±21.32</td>
<td>3292.01±1943.82</td>
<td>0.47±0.00</td>
<td>0.62±0.00</td>
<td>0.53±0.00</td>
<td>59.83±0.4</td>
</tr>
<tr>
<td>BIC-aiNet</td>
<td>640.57±22.26</td>
<td>43.94±20.07</td>
<td>3962.22±1517.24</td>
<td>0.49±0.00</td>
<td>0.60±0.00</td>
<td>0.51±0.00</td>
<td>80.58±4.7</td>
</tr>
<tr>
<td>SSB</td>
<td>170.52±39.65</td>
<td>72.69±22.55</td>
<td>1643.94±3541.26</td>
<td>0.46±0.00</td>
<td>0.63±0.00</td>
<td>0.61±0.00</td>
<td>34.61±31.69</td>
</tr>
<tr>
<td>SEBI</td>
<td>279.81±24.14</td>
<td>67.64±13.01</td>
<td>3221.63±208.18</td>
<td>0.47±0.00</td>
<td>0.61±0.00</td>
<td>0.53±0.00</td>
<td>97.62±2.4</td>
</tr>
</tbody>
</table>

4.2 Biological validation

For analyzing the biological significance of the algorithms, the proportions of the enriched biclusters by gene set enrichment analysis (GSEA) using GO and the weight enrichment score are adopted in this paper (Ashbumer et al., 2000). A bicluster is significantly enriched if p-values of GO terms annotated by GSEA are smaller than the predefined significance level. Then the proportion of significantly enriched biclusters can evaluate the quality of the biclustering algorithm. The higher its proportion is, the better the biclustering algorithm (Prelic et al., 2006; Nepormuceno et al., 2011). However, since the proportion of the enriched biclusters is very dependent on the gene number of the bicluster and it cannot quantitatively evaluate the quality of the bicluster, the Weight Enrichment score (WEScore) is also adopted (Li et al., 2012). The WEScore of a bicluster $B(i,j)$ is expressed as:

$$WEScore = \sum_{i=1}^{n} x_i s_i / r_j$$  \hspace{1cm} (15)
In (15) \( n \) is the number of GO terms obtained by GSEA using GO, \( x_i \) and \( s_i \) are the gene number and -log10 transformed p-value of \( i \)th GO terms. It can be seen from (15) that WEScore is independent with regard to the gene number of a bicluster, so WEScore can quantitatively measure the biological significance of a bicluster. The larger WEScore of the bicluster is, the biological significance of a bicluster is higher. This process of biological validation initially performs GSEA by the package of GOstats (Falcon and Gentleman, 2007), then uses the hyper geometric tests for computing p-value, and finally uses Benjamin-Hochberg False Discovery Rate (FDR) for performing multiple testing corrections (Purvesh and Sorin, 2005). The eight significant levels used in this work are 0.001%, 0.005%, 0.01%, 0.05%, 0.1%, 0.5%, 1% and 5%. According to the above process, the proportion of significantly enriched biclusters in different significant levels and WEScore as well as its mean and standard variance are computed for each bicluster. Additionally, the biological validations of the algorithms are based on the commonly used data of the Yeast Cycle data and the Gasch Yeast data.

Table 5 presents the aggregated proportion of the enriched biclusters in the eight significant levels and WEScore of each algorithm on the above two data, its first column showing the type and the name of the biclustering algorithms which is the same as the above Table 3, the next eight columns show the mean proportion of the enriched biclusters under eight significant levels, with the last column showing the mean WEScore. Table 6 gives the aggregated result of the five stochastic searching algorithms using four evaluating criteria. In addition, to facilitate the subsequent analysis, the eight significant levels are classified as low significance levels of 0.001%, 0.005% and 0.01%, middle significance levels of 0.05% and 0.1% and high significance levels of 0.5%, 1% and 5%.

As shown in Table 5, the mean proportion of the enriched biclusters produced by the FLOC of systematic algorithm is the highest, next is the stochastic algorithms using MSR while that using VE is the lowest. This is because the gene number of the bicluster produced by FLOC and stochastic algorithms using MSR is far greater than that using other cases as shown Table 3 and there is a significantly positive correlation between the proportion of the enriched biclusters and the gene number of a bicluster.

Table 5 The aggregated proportion of the enriched biclusters and WEScore on two yeast data

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>p&lt;0.001%</th>
<th>p&lt;0.005%</th>
<th>p&lt;0.01%</th>
<th>p&lt;0.05%</th>
<th>p&lt;0.1%</th>
<th>p&lt;0.5%</th>
<th>p&lt;1%</th>
<th>p&lt;5%</th>
<th>WEScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLOC</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>100.00%</td>
<td>2.02</td>
</tr>
<tr>
<td>ISA</td>
<td>47.50%</td>
<td>49.50%</td>
<td>51.50%</td>
<td>87.50%</td>
<td>90.50%</td>
<td>93.00%</td>
<td>95.00%</td>
<td>96.00%</td>
<td>5.23</td>
</tr>
<tr>
<td>OPSM</td>
<td>53.85%</td>
<td>61.54%</td>
<td>65.39%</td>
<td>73.08%</td>
<td>80.77%</td>
<td>84.62%</td>
<td>84.62%</td>
<td>88.46%</td>
<td>7.14</td>
</tr>
<tr>
<td>WEScore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In addition, Table 5 shows the interesting phenomena that although the proportion of the enriched biclusters produced by the stochastic searching algorithm is relatively small under low significance levels, its antecedent rises rapidly with a rise in the significance level. Furthermore, the proportion of the enriched biclusters rapidly rises under the significance level from p<0.01% to p<0.05 and from p<0.1% to 0.5%. Taking VE as an example, it is 53.40% under p<0.01% but rapidly rises to 61.50% under p<0.05%, and it is 67.60% under p<0.1% but rapidly rise to 84.90%under p<0.5%. This is why the eight significance levels are classified as the above. However, the proportions of the enriched biclusters under a high significance level are almost the same which prevents it from discriminating the biological significance of the algorithm. So WEScore appears essential in the above cases. The last column of Table 5 shows that although the mean proportion of the enriched biclusters produced by systematic algorithm using VE is the second best, its WEScore is the best, followed by stochastic searching algorithm using ASR and ACV, while the WEScore of FLOC while is the best in the he proportion of the enriched biclusters is trial.

Table 6 The aggregated proportion of the enriched biclusters and WEScore of five stochastic algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>p&lt;0.00 1%</th>
<th>p&lt;0.00 5%</th>
<th>p&lt;0.01 %</th>
<th>p&lt;0.05 %</th>
<th>p&lt;0.1 %</th>
<th>p&lt;0.5 %</th>
<th>p&lt;1%</th>
<th>p&lt;5 %</th>
<th>WEScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAB</td>
<td>64.00%</td>
<td>68.13%</td>
<td>70.88%</td>
<td>78.50%</td>
<td>83.75%</td>
<td>96.88%</td>
<td>99.63%</td>
<td>100.00%</td>
<td>9.52</td>
</tr>
<tr>
<td>PSOB</td>
<td>52.00%</td>
<td>55.75%</td>
<td>57.25%</td>
<td>66.75%</td>
<td>72.50%</td>
<td>93.00%</td>
<td>97.88%</td>
<td>99.75%</td>
<td>11.27</td>
</tr>
<tr>
<td>BIC-aiNet</td>
<td>58.00%</td>
<td>62.38%</td>
<td>65.13%</td>
<td>70.75%</td>
<td>74.25%</td>
<td>87.63%</td>
<td>90.88%</td>
<td>95.25%</td>
<td>9.56</td>
</tr>
<tr>
<td>SSB</td>
<td>48.75%</td>
<td>52.63%</td>
<td>54.50%</td>
<td>65.25%</td>
<td>71.50%</td>
<td>86.63%</td>
<td>91.88%</td>
<td>98.63%</td>
<td>13.78</td>
</tr>
<tr>
<td>SEBI</td>
<td>52.50%</td>
<td>55.50%</td>
<td>57.88%</td>
<td>68.75%</td>
<td>75.25%</td>
<td>94.88%</td>
<td>98.63%</td>
<td>99.88%</td>
<td>13.18</td>
</tr>
</tbody>
</table>
Table 6 summarizes the aggregate result of the five stochastic searching algorithms. Its first column shows the name of the stochastic searching algorithm and the other columns as being the same as that of Table 4. As Table 6 shows, the mean proportion of the enriched biclusters and WEScore of SAB are the best while those of SSB are the worst. In contrast, SSB is the best in the WEScore while SAB is the best.

5. DISCUSSION AND CONCLUSIONS

By analyzing performance criteria and the biological significance of the biclusters produced by the chose biclustering algorithms, the following interesting results are observed.

1. Various algorithms are suitable for various data, such as SSB being suitable for GASCH yeast expression data, while SAB better suits Yeast Cycle expression data.

2. Even with the same algorithm various coherence criteria are suitable for various data, such as MSR combined with stochastic searching algorithms being suitable for Yeast Cycle expression data, while VE is better suited for GASCH yeast expression data.

3. In terms of time performance the systematic biclustering algorithms are better than the stochastic biclustering algorithms.

4. Among four coherence criteria, MSR focuses on the biclusters with constant values, while VE, ACV, and ASR reflect the similarity of gene expression level and are suitable for the biclusters with a scaling pattern and the biclusters with coherent evolutions. In addition, the stochastic searching algorithms using VE and ACV as coherence criteria can find the more meaningful biclusters.

5. The threshold selections of coherence criterion depend on the data with prior knowledge. In addition, the threshold selections of MSR depend on the value range of the data; that is, the threshold selection plays a significant role in searching the significant biclusters.

6. As the biclusters produced by different biclustering algorithms in the same data appear to be very different even using the same threshold, a different algorithm should try a different threshold for obtaining better biclusters.

7. Considering that the biclusters obtained by the same biclustering algorithm using different coherence criteria are very different while the bicluster obtained by different biclustering algorithms using the same coherence criteria are similar, selecting appropriate coherence criterion is more important than selecting a biclustering algorithm.

8. The gene number of systematic biclustering algorithms is smaller than those of stochastic biclustering algorithms, so the systematic biclustering algorithm is suitable for expression data from a microscopic view while the stochastic biclustering algorithm is suitable for the expression data from a macroscopic view.
In total, the biclusters of 11 biclustering algorithms on three gene expression data are analyzed according to the 7 evaluation indices and biological significance in this work. There is no literature to which the multi-objective biclustering algorithm can be compared, so its performance will have to be analyzed in future work.

6. ACKNOWLEDGEMENTS

This work was supported by School of Computer Science and Engineering, University of Electronic Science and Technology of China, and school of computer and software, Huaiyin Institute of Technology. This research was supported in part by the National Natural Science Foundation of China (NSFC) under grants 60903074 and the National High Technology Research and Development Program of China (863 Program) under grant 2008AA01Z119.

7. REFERENCE


Cheng Y., Church G.M. (2000). Biclustering of expression data, Proceeding of the eighth international conference on intelligent system for molecular biology, California, USA.


